IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

# Evaluation of KNN & Random Forest in Prediction of Diabetes

**Divyanshu Singh, Gaurav Singh, Naincy Srivastava, Harsit Singh**
Department of Computer Science and Engineering (Data Science)
Raj Kumar Goel Institute of Technology, Ghaziabad.
contactdivyanshurajput@gmail.com 1011gauravsingh@gmail.com
naincy0807@gmail.com, harsitsingh17@gmail.com

**Abstract:** *Diabetes has actually become a prevalent health and wellness issue, with substantial effects for worldwide death prices. Every couple of mins a specific catch suddenly to diabetes- related issues. Scientist worldwide are working together with doctor to boost diabetic issues medical diagnosis as well as monitoring. Nonetheless the combination of artificial intelligence strategies holds assurance in enhancing analysis procedures along with lowering the dependence on substantial screening. This paper intends to forecast the start of diabetic issues in people well in advance consequently decreasing the demand for pricey therapies along with drugs. Leveraging advanced artificial intelligence techniques such as Random Forest and also K-nearest next-door neighbors (KNN), this research discovers anticipating versions for diabetic issues discovery. These formulas use affordable choices to typical analysis techniques helping with very early treatment and also monitoring. Accentuating a dataset sourced from reliable systems like Kaggle, making up details from 768 people the research assesses the precision of the KNN as well as Random Forest formulas. Relative evaluation versus existing research study highlights the efficiency of these formulas in diabetic issues discovery. Inevitably the application of such anticipating versions holds the possible to minimize diabetes-related death prices on a worldwide range.*

**Keywords:** KNN, Random Forest, Diabetes, Artificial Neural Networks, Logistic Regression

## I. INTRODUCTION

Diabetic issues stand as one of one of the most dangerous and also serious illness worldwide. This metabolic disorder interrupts the body's capability to control blood glucose degrees successfully causing a myriad of problems that can be deadly. Its occurrence covers throughout every edge of the world influencing people no matter geographical place or socioeconomic standing. Alarmingly, lots of people stay uninformed of their diabetic person standing up until signs and symptoms reveal or issues emerge. The signs of diabetic issues include a variety of indications, consisting of tiredness, lightheadedness, constant peeing, inexplicable weight reduction, as well as obscured vision. These indications work as important signs of the illness's existence, triggering people to look for clinical focus for medical diagnosis and also management. Handling diabetic issues positions substantial obstacles primarily as a result of the inflated prices connected with therapy and also treatment. Accessibility to detailed diabetes mellitus treatment continues to be Limited, leaving many people not able to manage necessary drugs, tracking products along with specialized clinical treatments. As a result, the death price amongst people with diabetic issues stays alarmingly high especially in areas with minimal medical care framework together with sources. Very early discovery of diabetic issues is critical in minimizing the threat of difficulties as well as sudden death. Timely medical diagnosis makes it possible for healthcare suppliers to launch suitable treatments consisting of way of life adjustments, medicine treatment and also routine tracking. According to stats from the World Health Organization, about 463 million grownups were coping with diabetic issues in 2019. Sadly, people identified with diabetic issues deal

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

with a dramatically minimized life span, with numerous succumbing to issues within a couple of years of medical diagnosis particularly if they do not have accessibility to sufficient healthcare solutions.

In the context of diabetic issues discovery, utilizing artificial intelligence formulas provides an affordable method to determining people in danger plus enhancing analysis precision. Formulas such as choice trees, K Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Logistic Regression (LR), to name a few have actually revealed assurance in discovering diabetic issues. These formulas have actually played a critical function in minimizing diabetes-related death prices by helping with very early discovery and also intervention. Despite the variety of artificial intelligence formulas offered, focus is put Random Forest along with K Nearest Neighbors (KNN) for their effectiveness in diabetic issues discovery. These formulas hold substantial capacity in aiding clients in determining diabetic issues and also starting prompt treatments to handle the problem efficiently. The dataset used in this research is sourced from trusted systems like Kaggle, which have actually been thoroughly made use of by scientists to check out numerous artificial intelligence comes close to for diabetic issues discovery. In this research the Random Forest as well as K Nearest Neighbor formulas are contrasted to establish their anticipating efficiency in spotting diabetic issues. The covering objective is to determine one of the most exact formula for anticipating diabetic issues along with boosting person end results. The preliminary action includes visualizing the dataset to determine elements adding to diabetes. This exploratory evaluation help in recognizing the dataset's features along with determining pertinent functions that affect diabetic issues danger. By leveraging artificial intelligence formulas doctor can improve the analysis procedure, boost precision as well as eventually decrease the worry of diabetic issues on people as well as medical care systems.

2nd action is information pre-processing where the information has actually been split right into 2 components i.e training component as well as the screening component. 70% of the information is made use of for educating the dataset as well as the 30% of the information is utilized to evaluate the dataset. Complying with that we ran each of our formulas independently to establish exactly how well they can anticipate our information. All the calculations were tried out the python making use of the system that has the setup Intel Core i5-11th generation@ 2.40 GHz laptop computer.

## II. LITERATURE REVIEW

Diabetes, a common persistent problem impacting millions internationally, needs very early recognition plus exact forecast for efficient monitoring along with problem avoidance. Lately the health care landscape has actually seen an rise in the fostering of machine learning (ML) strategies, providing appealing opportunities for illness forecast. The main factor of point of this literature review is to evaluate 2 noticeable ML formulas specifically K-Nearest Neighbors (KNN) as well as Random Forest within the context of diabetes mellitus forecast. Taking a look at this subject is extremely important as well as important for numerous factors.

Firstly exact diabetes mellitus forecast allows health care experts to step in proactively supplying customized treatment as well as endangering possible issues. ML versions experienced at identifying high-risk people can boost source appropriation and also support person results. Additionally the relative evaluation of KNN as well as Random Forest holds value. Both formulas are thoroughly used for category jobs and also realizing their efficiency in diabetes mellitus forecast is essential for specialists looking for to use one of the most appropriate formula in real life circumstances.

The research carried out by Devi Lal as well as Aswathy V. S. in 2023 ends that crossbreed artificial intelligence strategies, like the Hybrid Random Forest Algorithm, are most efficient for anticipating diabetes mellitus, accomplishing 90% precision. This study is a substantial action in boosting diabetes mellitus discovery plus avoidance, showcasing the relevance of choosing the appropriate formula for precise forecasts.

The research study, performed by Shamriz Nahzat together with Mete Yağanoğlu in 2021, wraps up that artificial intelligence especially Random Forest (RF) plus Artificial Neural Network (ANN) formulas, reveals assurance in reinventing diabetes mellitus threat forecast as well as very early recognition. Their study effectively recognized RF as

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

one of the most reliable classifier attaining an 88.31% precision price. This highlights the importance of very early diabetes mellitus discovery for efficient administration.

The research study by Dr. B. Premamayudu, K. Muralikrishna, along with K. Pramodh meant to evaluate the efficiency of the KNN classifier formula for diabetes mellitus forecast. Carried out at Vignan's Foundation for Science, Technology & & Research in India they accomplished a 79% precision price utilizing the scikit-learn collection in Python. This recommends capacity for utilizing KNN in health care for very early diabetes mellitus medical diagnosis profiting both individuals as well as health center monitoring with fast outcomes.

Led by Shahid Mohammad Ganie and also Majid Bashir Malik in 2022 this research contrasts artificial intelligence formulas for very early forecast of type-II diabetes mellitus. They boosted information established top quality with preprocessing methods plus recommended a block representation for ML design growth. Random Forest (RF) revealed the highest possible precision at 99.67% for the training collection coupled with 93.79% for the screening collection. The research study recommends discovering crossbreed or set techniques for far better outcomes as well as highlights the requirement for bigger datasets to boost diabetes mellitus forecast.

This research by Cindy Nabila Noviyanti together with Alamsyah in 2024 concentrates on very early diabetes mellitus discovery utilizing the Random Forest algorithm. They accomplished a precision of 87%, showcasing its prevalence over various other techniques with the exact same information. Nevertheless, additional research study is required to improve discovery precision by discovering variables like information harmonizing approaches, attribute choice, coupled with bigger datasets.

## III. USING RANDOM FOREST FOR PREDICTION

Among the well-known members of ensemble learning algorithms, Random Forest stands out. This is an algorithm that is similar to decision trees that are a crucial part of our proposed solution for solving classification and regression problems. The Random Forest follows a process known as bootstrapping, which first creates an ensemble of decision trees trained on distinct subsets of the dataset. These trees in the forest independently grow while giving class labels or values to input instances.

Random Forest works in a way that is analogous to decision tree only using different process. It produces more accurate results with better robustness than one tree does by combining predictions from several decision trees instead of relying on just a single one. Each Decision Tree in Random Forest uses only some randomly selected attributes at each node for splitting thus creating deviation across the component trees. Generalization can be improved by this reduction in diversity that overfits models.

Per instance, during the prediction stage, an input traverses each random forest decision tree whose final prediction is obtained through averaging in case of regression or majority vote for classification by all predictions from these trees. As a result of this ensemble approach, these predictions become more accurate and reliable thereby reducing bias and volatility typical for single decision trees. These words are important because they sum up the main ideas involved in Random Forest such as “bagging,” “majority voting” or “averaging”, “ensemble learning,” and “feature subset selection.” In other words, it is within these concepts that Random Forest modeling essential ideas and procedures are covered.

In our research paper, we illustrate how Random Forest can be used effectively. Let us see a situation where a business would want to predict customer attrition. By analyzing different consumer information (e.g. demographic features, past buying records) about its customers, a company may employ random forest to estimate the probability of customer churn occurring in any given period. This predictive capacity empowers corporations to develop focused retention plans that diminish churn enhancing customer loyalty.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

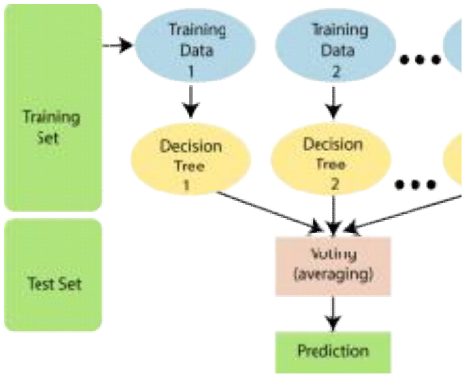


Figure 1- Flowchart of Random Forest

## IV. USING KNN IN PREDICTION

KNN's foundation is supervised learning. As illustrated by KNN, the most common application of this approach is usually in classification cases. It also relies on regression and classification for that matter.

The non-parametric nature of this algorithm simply means that it assumes nothing about the data underneath it. Also known as a "lazy learner," this algorithm does not begin to learn from its training set straightaway. Instead, after storing it during the classification phase, it applies the action on the dataset. On the other hand, there are some advantages of using this method such as easy implementation. The greatest robustness against noisy training data is attributed to this approach while large training data sets can serve as a better fit.

However, there are a couple problems with this technique.

For instance, ifk were unknown all through then it will be a difficult question since it always needs to be known. Moreover, computation is expensive in that it has to compute distances between each data point among the training samples. Let us assume we have to check if the image that we have

resembles a dog or a cat. Therefore, our problem will be solved by using KNN algorithm. This means that given vast majority of features from this new dataset such as similar features present in dog and cat pictures, it will assign dog or cat to each case.

## V. METHODOLOGY

**KNN- KNearest Neighbor methodology**

A. Dataset: The data has been taken from Kaggle website for prediction of diabetes which has also used by other researchers. The dataset was obtained freely on the internet

for public use and research purposes.

This is adf data frame with zero null values on its 768 patients' records showing different factors that may lead to diabetes. Number of diabetics This model predicts how many numbers of patients having diabetes are there. In the below output we observe 500 classes belong to label 0 and 268 classes belong to label 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

| S. No | Attributes |
|---|---|
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI (body mass index) |
| 7 | Diabetes pedigree function |
| 8 | age |

Table 1 – data attributes of patients

The 9thattribute is the class variable of each data points. This class variable shows the outcomes 1 & 0 for diabetes which indicates diabetic & non-diabetic.

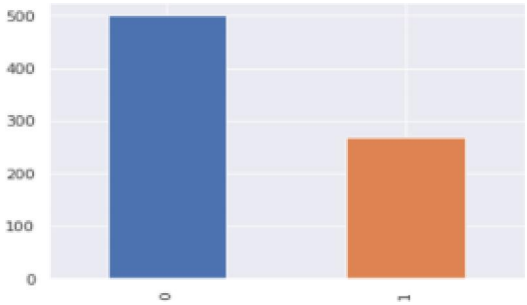


Figure 2 - Distribution of Diabetic and Non-Diabetic Cases Based on Class Variable

The above graph shows that the data is biased towards data points having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

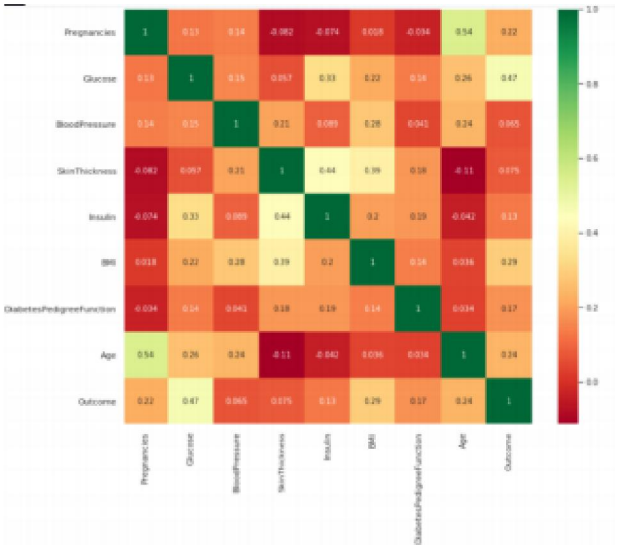


Figure 3 - correlation heatmap

Above values are the analytical worths of the dataset which we have actually utilized. Right here, from this connection matrix we familiarized that pregnancies plus glucose columns are really crucial to anticipate the outcome. These 2 columns played crucial function.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

```
plt.figure(figsize=(12,10))  # on this line I just set the size of figure to 12 by 10.
p=sns.heatmap(diabetes_data.corr(), annot=True,cmap ='RdYlGn')  # seaborn has very simple solution
for heatmap
```

**B. Data pre-processing:** This is one of the most important procedure. Primarily, medical care associated information might include numerous missing out on worths along with numerous errors which may create for reduced reliable of data. So to enhance the top quality coupled with efficiency information handling ought to be done. This procedure is extra necessary to obtain excellent precision. There are generally 2 enter this information pre-processing they are -.

1. Missing values removal.
2. Splitting of data into training and testing sets.

**C. Applying classifier strategy:** After training along with screening datasets are divided together with without null values in the dataset, we can currently use the machine learning classifier technique to the dataset. We have numerous category strategies such as support vector machine (SVM) however we are making use of K nearest neighbor's classifier technique. .

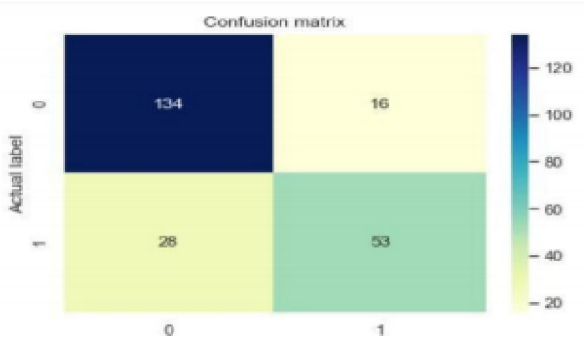


Figure 4 – Confusion Matrix

This approach This strategy utilized K- Nearest Neighbors (KNN) classifier method. Utilizing this artificial intelligence strategy, we discover the precision of forecasting diabetic issues making use of KNN formula. As well as we have actually obtained the precision rating of 81% which is much better to get prediction.

**Random Forest methodology**

The Random Forest methodology employed in this study is highlighted via procedural flowchart beginning with the acquisition of data, followed by preprocessing, dataset partitioning, model construction and concluding with performance assessment.

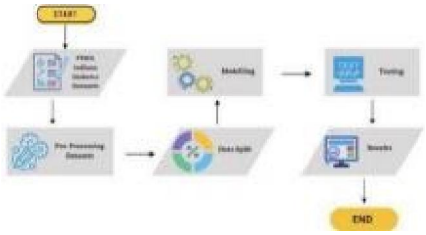


Figure 5 - Random Forest methodology

**A. Data Collection:**

The dataset utilized for anticipating diabetes mellitus was sourced from Kaggle a system typically made use of by researchers. This dataset, consisting of information of 768 people, is openly available and also lacking any type

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

of missing out on worths. It includes different elements possibly pertaining to diabetic issues. Notably the circulation of diabetic person clients exposes 500 circumstances classified '0' plus 268 circumstances classified '1'.

B. Preprocessing Data: Information pre-processing plays an important function in preparing information for modeling. Throughout this stage the concentrate gets on determining plus solving any kind of missing out on values (NaN) or vacant rows in the information collection. After performing this check it is validated that the information collection is full together with does not include any kind of NaN values. Additionally, an extensive assessment is executed to discover no values in each function. While particular functions, like the variety of pregnancies might practically have absolutely no values, it is impractical for various other attributes such as sugar focus, high blood pressure, skin density, insulin as well as BMI to be zero. In order to resolve this, the absolutely no worths in these attributes are changed with the mode value. C. Split Data: The dataset is separated right into 3 collections: training data (60%), validation data (25%), and testing data (20%). Training data is utilized to train the Random Forest model, while validation data is working to evaluate model performance during the training phase. Subsequently, testing data is used to judge the model's performance after training.

**D. Evaluation:** The efficiency analysis of the version is accomplished by using the confusion matrix which is a frequently made use of analysis method in set discovering designs. By contrasting the model's predictions to the actual outcomes, the confusion matrix offers valuable insights into the accuracy of the model. It consists of four essential elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From the confusion matrix, various metrics have arrived such as accuracy, precision, recall, and F1 scores, which are providing a comprehensive evaluation ofthe model's performance. These metrics given are shown in detail to estimate the effectiveness of the model in predicting the diabetes disease.

- Accuracy

Accuracy measures the accuracy of the model is in predicting the entire data. It is measured by the formula:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- Precision

Precision tells the mark to which the work predicted to be duplicate is fake It is measured by the formula:

$$Precision = \frac{TP}{(TP + FP)}$$

- Recall (Sensitivity)

Recall measures the level to which the model efficaciously detects fake jobs as a whole. It is measured by the formula:

$$Recall = \frac{TP}{(TP + FN)}$$

- F1-Score

The F1-Score is a grouping of precision and recall into a single metric that yields the total model performance. It is measured by the formula:

$$F1 - Score = \frac{2 * (Presisi * Recall)}{(Presisi + Recall)}$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

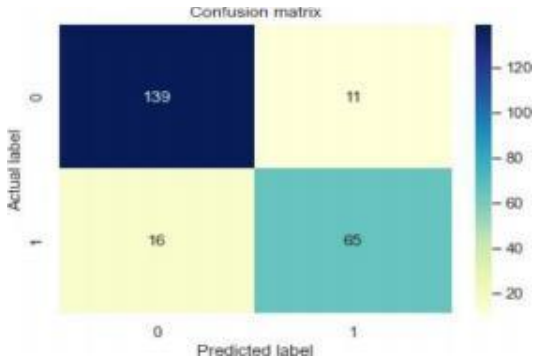


Figure 6 - Confusion matrix

In our study, the Random Forest algorithm was working to predict diabetes, achieving an accuracy of 88%. This high accuracy highlights the effectiveness of the Random Forest approach in the = diabetes detection. Similarly covers about one-third of the dataset.

## VI. EXPERIMENTAL RESULTS

| Classification Technique | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Random Forest | 88% | 82% | 89% | 91% |
| K's Nearest Neighbour | 81% | 89% | 83% | 86% |

Table 2 - Experimental Results

The Random Forest algorithm is known or its durability and also capacity to take care of huge datasets with high dimensionality. It functions by building numerous decison trees throughout training and also outputting the course that is the setting of the courses of the individual trees. This set method commonly cause high precision because of the balancing out result and also decreases overfitting.

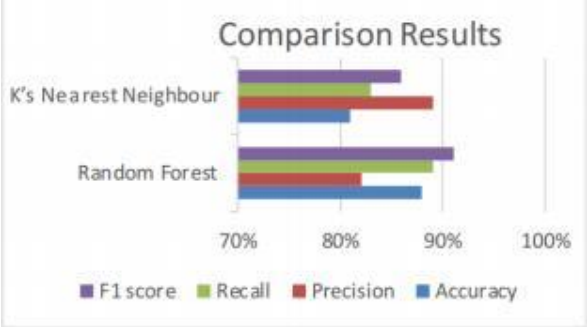


Figure 7 - Comparison Results

Beyond, the K-Nearest Neighbors (KNN) formula is a non- parametric technique utilized for category plus regression jobs. It classifies an information factor based upon just how its closest next-door neighbors are identified. KNN is easy to apply and also has the benefit of having the ability to adjust to modifications in the dataset. Nonetheless it can be computationally costly particularly with huge datasets as it calls for determining ranges in between the data factors.

The selection in between these algorithms relies on particular variables such as the features of the dataset computational sources plus the preferred degree of precision. If accuracy is important the Random Forest algorithms might be chosen as a result of its greater precision score.Additionally Random Forest often tends to take care of noisy data well and also is much less conscious outliers.

Alternatively, the KNN formula provides an extra nuanced strategy, which is specifically useful in instances where interpretability and also insurance coverage are vital. KNN's capacity to adjust to the information and also manage

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

nonlinear partnerships can be useful in specific scenarios specifically when managing smaller sized datasets or when interpretability is vital.

Eventually, the decision between these two algorithms should be based on the unique requirements and limitations of the problem at finger. Cautious factor to consider of elements such as dataset dimension, computational sources, interpretability, as well as preferred precision degree is necessary in picking one of the most ideal formula for diabetes predication or any kind of various other healthcare application.

## VII. CONCLUSION

In contrast, although both the Random Forest as well as K- Nearest Neighbors (KNN) formulas work in forecasting wellness problems like diabetes mellitus they vary in regards to precision plus insurance coverage. The Random Forest formula attains a precision of 88% showing its effectiveness in finding diabetes mellitus. Beyond KNN attains a somewhat reduced precision rating of 81%. Nonetheless KNN's accuracy as well as precision still make it a solid rival specifically considering its capacity to cover about one-third of the dataset. The option in between these formulas relies on certain aspects such as the attributes of the dataset, computational sources as well as the preferred degree of precision If accuracy is of utmost value, the Random Forest algorithm may be preferred because of its higher accuracy rate. On the other hand, the KNN formula supplies a much more nuanced method which is especially beneficial in situations where interpretability as well as coverage are vital. Ultimately, the decision between these two algorithms should be based on the unique requirements and limitations of the problem at hand.

## VIII. REFERENCES

[1] Lal, D., & Aswathy, V. S. (2023). Exploring the Efficacy of Machine Learning Algorithms for Diabetes Prediction: A Comparative Prediction. https://doi.org/10.22214/ijraset.2023.51565

[2] Nahzat, S., & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. European Journal of Science and Technology, (24), 53-59. DOI: 10.31590/ejosat.899716

[3] Premamayudu, B., Muralikrishna,K., & Pramodh, K. (2022). Diabetes Prediction Using Machine Learning KNN-Algorithm Technique. International Journal of Innovative Science and Research Technology, 7(5), 941. ISSN: 2456-2165. Retrieved fromwww.ijisrt.com

[4] Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. Journal of Information Systems Engineering and Radio Science, 2(1). https://doi.org/10.52465/joiser.v2i1.245

[5] Ganie, S. M., & Malik, M. B. (2022). Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus. International Journal of Medical Engineering and Informatics, 14(6), 473–483

[6] Barakat,N. H., Bradley, A. P., & Barakat, M. N. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in Biomedicine, 14(4), 1114- 1120. DOI: 10.1109/TITB.2009.2039485

[7] Hasan, M. K., Alam, M. A., Das, D., & Hossain, E. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, PP(99), 1-1. DOI: 10.1109/ACCESS.2020.298985

[8] Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence (IJ-AI), 11(1), 284. DOI: 10.11591/ijai.v11.i1.pp284-290

[9] Amulya, K. J., Divya, S., Deepali, H. V., & Ravikumar, V. (2020). A Survey on Diabetes Prediction Using Machine Learning. In ICCCE 2020, Proceedings of the 3rd International Conference on Communications and Cyber Physical Engineering (pp. 1049-1057). DOI: 10.1007/978-981-15-7961-5_

[10] Kaggle. Predict Diabetes Dataset. Retrieved from https://www.kaggle.com/datasets/whenamancodes/pre dict-diabities

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

[11] Fatima Afzaal. (2023). Diabetes Prediction Project Using Random Forest [Jupyter Notebook]. Retrieved from GitHub: https://github.com/fatimaAfzaal/Diabetes-Prediction Project-Using-Random Forest/blob/main/Diabetes_Prediction_Project_Using _Random_Forest.ipynb
[12] Pradnya1208. (2021). Diabetes classification using KNN [Python code]. Retrieved from GitHub: https://github.com/Pradnya1208/Diabetes classification-using-KNN